yang_seonhyeHW21

```
library(mfp)
```

```
## Loading required package: survival
```

```
data("bodyfat")
```

Question 0

```
# Doing a 50-50 Training-Testing split
set.seed(10)
sampleRows <- sample.int(nrow(bodyfat), floor(.5 * nrow(bodyfat)))
training <- bodyfat[sampleRows, ]
testing  <- bodyfat[-sampleRows, ]
```

Question 1

Part A: Full model using all predictors

```
fullModel <- lm(brozek~. - siri - density - case, data = training)
summary(fullModel)
```

```
##
## Call:
## lm(formula = brozek ~ . - siri - density - case, data = training)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.8951 -2.8066 -0.2579  2.9734  8.0830
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -41.17559   27.59768  -1.492   0.1385
## age           0.09630    0.04480   2.150   0.0337 *
## weight       -0.11116    0.07398  -1.503   0.1357
## height        0.17980    0.23077   0.779   0.4375
## neck         -0.32141    0.29362  -1.095   0.2760
## chest         0.06575    0.14716   0.447   0.6559
## abdomen       0.77675    0.12440   6.244 7.86e-09 ***
## hip          -0.17483    0.17913  -0.976   0.3312
## thigh         0.37162    0.18157   2.047   0.0430 *
## knee          0.05794    0.32716   0.177   0.8597
## ankle         0.13576    0.29284   0.464   0.6438
## biceps        0.21338    0.21410   0.997   0.3211
## forearm       0.34068    0.25955   1.313   0.1920
## wrist        -1.65580    0.70665  -2.343   0.0209 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.076 on 112 degrees of freedom
```

```
## Multiple R-squared:  0.7687, Adjusted R-squared:  0.7418
## F-statistic: 28.63 on 13 and 112 DF,  p-value: < 2.2e-16
```

Part B: Exhaustive search

```
library(leaps)
x <- model.matrix(brozek~. - 1 - siri - density - case, data = training)
y <- training$brozek
bestmods <- leaps(x, y, nbest = 1)
cols <- colnames(x)[which(bestmods$which[which.min(bestmods$Cp), ], arr.ind = TRUE)]
exhaustiveModel <- lm(as.formula(paste("brozek ~ ",paste(cols, collapse="+"),sep = "")), data = training
summary(exhaustiveModel)
```

```
##
## Call:
## lm(formula = as.formula(paste("brozek ~ ", paste(cols, collapse = "+"),
##     sep = "")), data = training)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.8957 -2.5755 -0.4325  2.9464  9.2570
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.29049    8.16523  -0.893  0.37374
## age          0.11125    0.04029   2.761  0.00669 **
## neck        -0.45481    0.25426  -1.789  0.07622 .
## abdomen      0.71412    0.08176   8.735 1.92e-14 ***
## hip         -0.31994    0.14475  -2.210  0.02901 *
## thigh        0.38532    0.15739   2.448  0.01583 *
## forearm      0.44235    0.23749   1.863  0.06501 .
## wrist       -1.72740    0.62914  -2.746  0.00698 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.019 on 118 degrees of freedom
## Multiple R-squared:  0.7631, Adjusted R-squared:  0.749
## F-statistic:  54.3 on 7 and 118 DF,  p-value: < 2.2e-16
```

Part C: Exhaustive search using Adjusted R Squared

```
x <- model.matrix(brozek~. - 1 - siri - density - case, data = training)
y <- training$brozek
bestmods <- leaps(x, y, nbest = 1, method = "adjr2")
cols <- colnames(x)[which(bestmods$which[which.max(bestmods$adjr2), ], arr.ind = TRUE)]
exhaustiveModelRSq <- lm(as.formula(paste("brozek ~ ",paste(cols, collapse="+"),sep = "")), data = trai
summary(exhaustiveModelRSq)
```

```
##
## Call:
## lm(formula = as.formula(paste("brozek ~ ", paste(cols, collapse = "+"),
##     sep = "")), data = training)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -9.8957 -2.5755 -0.4325  2.9464  9.2570
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.29049    8.16523  -0.893  0.37374
## age          0.11125    0.04029   2.761  0.00669 **
## neck        -0.45481    0.25426  -1.789  0.07622 .
## abdomen      0.71412    0.08176   8.735 1.92e-14 ***
## hip         -0.31994    0.14475  -2.210  0.02901 *
## thigh        0.38532    0.15739   2.448  0.01583 *
## forearm      0.44235    0.23749   1.863  0.06501 .
## wrist       -1.72740    0.62914  -2.746  0.00698 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.019 on 118 degrees of freedom
## Multiple R-squared:  0.7631, Adjusted R-squared:  0.749
## F-statistic:  54.3 on 7 and 118 DF,  p-value: < 2.2e-16
```

Part D: Forward Selection

```
library(MASS)
indep.vars <- ~age + weight + height + neck + chest + abdomen + hip + thigh + knee + ankle + biceps + f
object <- lm(brozek ~ 1, data = training)
aicModel <- step(object, scope = indep.vars, direction = "forward")
```

```
## Start:  AIC=525.69
## brozek ~ 1
##
##           Df Sum of Sq    RSS    AIC
## + abdomen  1    5628.8 2414.4 376.07
## + chest    1    4701.4 3341.9 417.03
## + weight   1    3836.3 4207.0 446.04
## + hip      1    3797.4 4245.9 447.19
## + thigh    1    3364.7 4678.5 459.42
## + biceps   1    2692.8 5350.4 476.33
## + knee     1    2559.3 5483.9 479.43
## + neck     1    2505.0 5538.3 480.68
## + forearm  1    1484.9 6558.4 501.98
## + wrist    1    1439.1 6604.2 502.86
## + ankle    1    1016.3 7026.9 510.67
## + age      1     419.7 7623.6 520.94
## <none>                 8043.2 525.69
## + height   1      14.2 8029.0 527.47
##
## Step:  AIC=376.07
## brozek ~ abdomen
##
##           Df Sum of Sq    RSS    AIC
## + weight   1   278.390 2136.0 362.63
## + wrist    1   183.995 2230.4 368.08
## + hip      1   180.692 2233.7 368.27
## + neck     1   145.964 2268.4 370.21
```

```
## + age      1    71.390 2343.0 374.29
## + height   1    68.003 2346.4 374.47
## + ankle    1    43.387 2371.0 375.78
## + knee     1    39.566 2374.8 375.99
## <none>                 2414.4 376.07
## + chest    1    33.504 2380.9 376.31
## + thigh    1    22.164 2392.2 376.91
## + biceps   1    17.442 2397.0 377.16
## + forearm  1     1.128 2413.3 378.01
##
## Step:  AIC=362.63
## brozek ~ abdomen + weight
##
##           Df Sum of Sq    RSS    AIC
## + thigh    1    42.008 2094.0 362.13
## + wrist    1    39.185 2096.8 362.30
## <none>                 2136.0 362.63
## + knee     1    32.358 2103.7 362.71
## + biceps   1    32.096 2103.9 362.73
## + forearm  1    20.297 2115.7 363.43
## + neck     1    13.705 2122.3 363.82
## + height   1     7.628 2128.4 364.18
## + hip      1     5.651 2130.4 364.30
## + age      1     3.820 2132.2 364.41
## + ankle    1     2.267 2133.8 364.50
## + chest    1     0.276 2135.8 364.62
##
## Step:  AIC=362.13
## brozek ~ abdomen + weight + thigh
##
##           Df Sum of Sq    RSS    AIC
## + hip      1    33.954 2060.1 362.07
## <none>                 2094.0 362.13
## + wrist    1    31.028 2063.0 362.25
## + age      1    25.464 2068.6 362.59
## + biceps   1    22.109 2071.9 362.79
## + height   1    17.108 2076.9 363.10
## + knee     1    16.461 2077.6 363.14
## + neck     1    15.271 2078.8 363.21
## + forearm  1    13.199 2080.8 363.33
## + chest    1     5.278 2088.7 363.81
## + ankle    1     1.479 2092.5 364.04
##
## Step:  AIC=362.07
## brozek ~ abdomen + weight + thigh + hip
##
##           Df Sum of Sq    RSS    AIC
## + wrist    1    34.253 2025.8 361.96
## <none>                 2060.1 362.07
## + neck     1    26.492 2033.6 362.44
## + age      1    21.380 2038.7 362.76
## + knee     1    17.520 2042.5 362.99
## + biceps   1    14.156 2045.9 363.20
## + forearm  1    10.511 2049.6 363.43
```

```
## + height   1     8.427 2051.6 363.55
## + chest    1     2.493 2057.6 363.92
## + ankle    1     1.336 2058.7 363.99
##
## Step:  AIC=361.96
## brozek ~ abdomen + weight + thigh + hip + wrist
##
##           Df Sum of Sq    RSS    AIC
## + age      1    59.473 1966.3 360.20
## <none>                  2025.8 361.96
## + biceps   1    29.641 1996.2 362.10
## + forearm  1    26.487 1999.3 362.30
## + knee     1    26.225 1999.6 362.32
## + neck     1    12.337 2013.5 363.19
## + ankle    1     7.428 2018.4 363.49
## + height   1     5.749 2020.1 363.60
## + chest    1     2.770 2023.0 363.79
##
## Step:  AIC=360.2
## brozek ~ abdomen + weight + thigh + hip + wrist + age
##
##           Df Sum of Sq    RSS    AIC
## + forearm  1    47.097 1919.2 359.15
## <none>                  1966.3 360.20
## + biceps   1    25.520 1940.8 360.56
## + neck     1    20.764 1945.6 360.87
## + knee     1    12.330 1954.0 361.41
## + height   1    10.472 1955.9 361.53
## + ankle    1     5.946 1960.4 361.82
## + chest    1     4.107 1962.2 361.94
##
## Step:  AIC=359.15
## brozek ~ abdomen + weight + thigh + hip + wrist + age + forearm
##
##          Df Sum of Sq    RSS    AIC
## <none>                 1919.2 359.15
## + neck    1   27.9705 1891.3 359.30
## + height  1   12.3747 1906.9 360.33
## + biceps  1   11.6976 1907.5 360.38
## + knee    1    7.3066 1911.9 360.67
## + ankle   1    6.5187 1912.7 360.72
## + chest   1    0.0140 1919.2 361.15
```

```
summary(aicModel)
```

```
##
## Call:
## lm(formula = brozek ~ abdomen + weight + thigh + hip + wrist +
##     age + forearm, data = training)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.8607 -2.4999 -0.3756  3.2781  7.8340
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -24.38655   12.60033  -1.935   0.0553 .
## abdomen       0.75623    0.09296   8.135 4.72e-13 ***
## weight       -0.07024    0.04596  -1.528   0.1291
## thigh         0.35856    0.15589   2.300   0.0232 *
## hip          -0.20683    0.16296  -1.269   0.2069
## wrist        -1.68627    0.65736  -2.565   0.0116 *
## age           0.09226    0.04158   2.219   0.0284 *
## forearm       0.40362    0.23719   1.702   0.0915 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.033 on 118 degrees of freedom
## Multiple R-squared:  0.7614, Adjusted R-squared:  0.7472
## F-statistic: 53.79 on 7 and 118 DF,  p-value: < 2.2e-16
```

Question 2

```
mseModel1 <- mean((testing$brozek - predict.lm(fullModel, testing)) ^ 2)
mseModel2 <- mean((testing$brozek - predict.lm(exhaustiveModel, testing)) ^ 2)
mseModel3 <- mean((testing$brozek - predict.lm(exhaustiveModelRSq, testing)) ^ 2)
mseModel4 <- mean((testing$brozek - predict.lm(aicModel, testing)) ^ 2)
print(mseModel1)
```

```
## [1] 17.15662
```

```
print(mseModel2)
```

```
## [1] 16.57527
```

```
print(mseModel3)
```

```
## [1] 16.57527
```

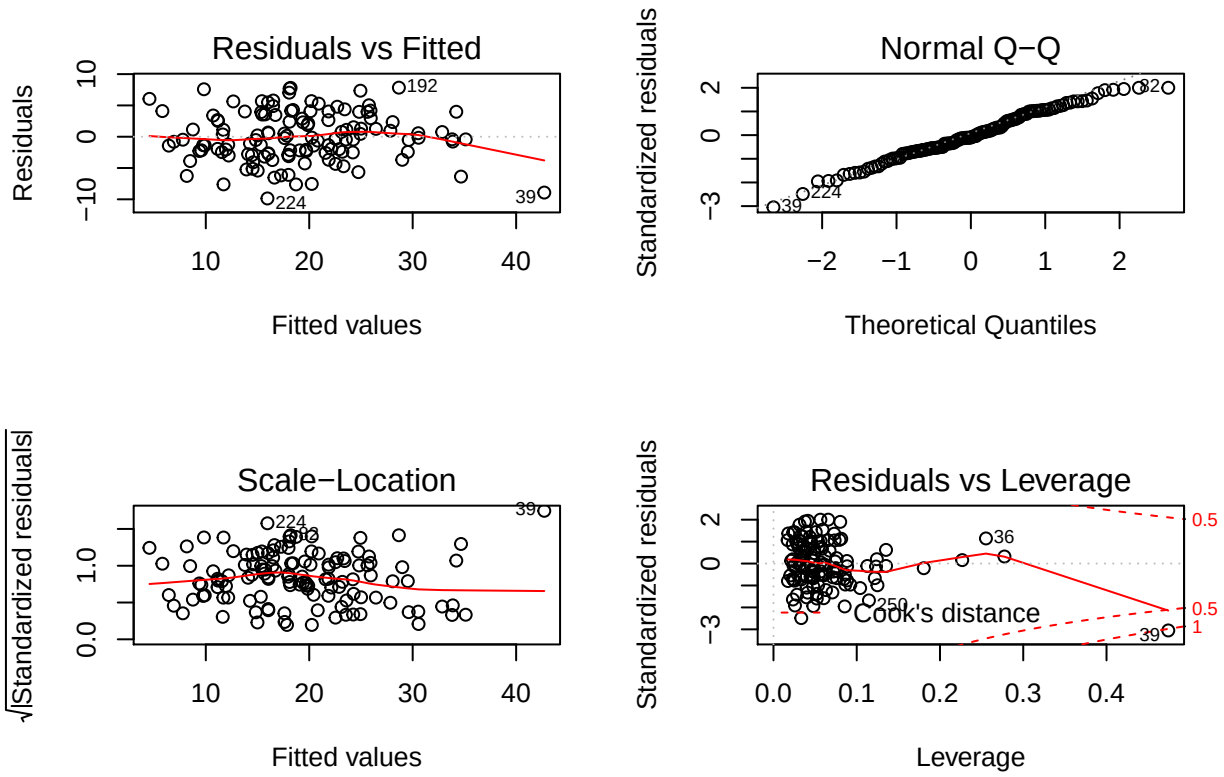
```
print(mseModel4)
```

```
## [1] 16.32268
```

Question 3

Looking at the MSE Values provided by question 2, we can see that the model with the lowest test MSE was model 4, or the AIC step model. Second was the exhaustive search R^2 model and the exhaustive C_p search which were equal, followed by the full model. We can intrepret the best model as follows:

```
par(mfrow=c(2,2))
plot(aicModel)
```



Looking at the Residuals Vs Fitted, we can see a flat trendline along with relatively equally spread residuals, indicating we have a linear relationship. Our Normal Q-Q plot is also extremely linear, indicating that residuals are normally distributed. The Scale-Location plot also has a nearly flat, linear trendline, along with residuals equally spread along predictors, so we can assume homoscedasticity. Finally, looking at our Residuals Vs Leverage, we can see that there is a relatively flat trendline with most of our points indicating very little to no leverage but one or two points that have high leverage, and should probably be removed from the dataset to have a better fit. Overall, we can see from these plots that Model 4, the forward step AIC model is a relatively good model, and can be used for decent predictions with this dataset.